PROJECT SYNOPSIS ON: Vital Sign based Disease Identification using Machine-learning algorithm

Final Report Submitted in Partial Fulfillment of the Requirements for the degree of

Bachelor of Technology in Applied Electronics and Instrumentation Engineering

By

SWETA KUMARI-13005316016
BASU VERMA-13005316060
ABHISHEK GANGULY-13005316069
ABHISHEK KUMAR-13005316068
ABHISHEK SINGH-13005316067

Under the guidance of
MRS. NABANITA CHAKRABORTY BANERJEE

Techno Main Salt Lake
EM 4/ Salt lake City,Sector V
Kolkata – 700091
2020

TECHNO INDIA

(WEST BENGAL UNIVERSITY OF TECHNOLOGY)

Faculty of AEIE Department

Certificate of Recommendation

This is to certify that Sweta Kumari (16), Basu Verma (60), Abhishek Singh (67), Abhishek Kumar (68) and Abhishek Ganguly (69) have been involved in their project work titled "**Vital Sign based Disease Identification using Machine-learning algorithm**", under the direct supervision and guidance of Mrs. Nabanita Chakraborty Banerjee. I am satisfied with their work, which is being presented for the partial fulfillment of the degree of Bachelor of Technology in Applied Electronics and Instrumentation Engineering, West Bengal University of technology (WBUT), Kolkata–700032.

Mrs.Nabanita Chakraborty Banerjee

"Name of Teacher"
(Teacher in charge of Project)

Date:

Prof. Sanghmitra Manna
HOD
Department of AEIE
(Techno India)

Date:16th June 2020

WEST BENGAL UNIVERSITY OF TECHNOLOGY

Faculty of AEIE Department

Provisional Certificate of Approval *

The foregoing project synopsis is hereby approved as a creditable study of Bachelor of Technology and presented in a manner satisfactory to warrant its acceptance as a pre-requisite to the final semester examination for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or any statement made, opinion expressed or conclusion therein but approve this project only for the purpose for which it is submitted.

Signature of the examiners

Abstract:

The relationship between stress and illness is complex. Stressors have a major influence upon mood, our sense of well-being, behavior, and health. Acute stress responses in young, healthy individuals may be adaptive and typically do not impose a health burden. However, if the threat is unremitting, particularly in older or unhealthy individuals, the long-term effects of stressors can damage health.

Earlier, we first took some pulse samples for a pulse sensor as well as ECG samples and calculated the heart rate of different persons. This was done by various detection methods like peak to peak detection as well as QRS analysis .On analysis of the result we found the clear difference between stressed and non-stressed person

Thus, our goal is to detect the stress affected person and identify and predict the diseases which a person can suffer through if the stress environment is not checked upon. One of the stress related disease is Arrhythmia. Arrhythmia is considered a life-threatening disease causing serious health issues in patients, when left untreated. An early diagnosis of arrhythmias would be helpful in saving lives. This study is conducted to classify patients into one of the sixteen subclasses, among which one class represents absence of disease and the other fifteen classes represent electrocardiogram records of various subtypes of arrhythmias. The research is carried out on the dataset taken from the University of California at Irvine Machine Learning Data Repository. The dataset contains a large volume of feature dimensions but we have selected only that feature which we have extracted from MATLAB like heat rate, QRS duration, P-R interval, Q-T interval and many more. In this study the different algorithm for classification in machine learning are used. A few popular techniques from contemporary literature were implemented namely k-nearest neighbors (**KNN**), Support vector classifier (SVC), RandomForests and Xgboost. We used K folds cross validation to detect over fitting. K-Fold is a popular and easy to understand, it generally results in a less biased model compare to other methods. Because it ensures that every observation from the original dataset has the chance of appearing in training and test set. This is one among the best approach if we have a limited input data**.** In doing so, we identify the learning methodologies utilized, data sources, appropriate means of model evaluation, and specific challenges of classifying the disease. This then leads us to propose whether a person suffering from a disease or not framework through which ML can be used as a learning strategy. Our research will hopefully be informative and of use to those performing future research in this application area. The accuracy of the model was found to be 70 percent.

Keywords: ECG samples, Arrhythmia, Random Forest classification, Support vector Machine (SVM), K-nearest neighbors (KNN), K-Folds Cross Validation

List of figures

Table of contents/Index:

Introduction:

According to Gary G. Berntson, J. Thomas Bigger, Jr., Dwain L. Eckberg, Paul Grossman, Peter G. Kaufmann, Marek Malik, Haikady N. Nagaraja, Stephen W. Porges, J. Philip Saul, Peter H. Stone, Maurits W. van der Molen Kline, Kronhaus, Moore, and Spear [32] did a research on stress and disease with ECG signal. “Modeling a stress signal” by Nandita Sharma, Tom Gedeon,2013 was very informative and helped a lot in getting the ideas of classification algorithm like SVC(Support vector classifier),KNN(K-Nearest Neighbor), Decision Trees and Random Forest and GA algorithm. We can use GA algorithm for feature selection and with the help of these algorithm we can build the model to classify whether the person is suffering from stressed disease or not.

With a society filled with competition, perfection, and performance, it is easy to get stuck in periods of stress. Stress is a defensive behavior from the body when danger, real or imaginary, is present. Unfortunately, stress damages the body in ways which usually are underestimated. It can have effects in many ways such as emotional, cognitive, physical and behavioral. Stress can be divided into mental and physical.

Studies have found many health problems related to stress. Stress seems to worsen or increase the risk of conditions like obesity, heart disease, Alzheimer's disease, diabetes, depression, gastrointestinal problems, and arrhythmia.

Irregularity in heart beat may be harmless or life threatening. Hence both accurate detection of presence as well as classification of arrhythmia are important. Arrhythmia can be diagnosed by measuring the heart activity using an instrument called ECG or electrocardiograph and then analyzing the recorded data. Different parameter values can be extracted from the ECG waveforms and can be used along with other information about the patient like age, medical history, etc. to detect arrhythmia. However, sometimes it may be difficult for a doctor to look at these long duration ECG recordings and find minute irregularities.

One of the common machine learning (ML) tasks, which involves predicting a target variable in previously unseen data, is classification. The aim of classification is to predict a target variable (class) by building a classification model based on a training dataset, and then utilizing that model to predict the value of the class of test data. This type of data processing is called supervised learning since the data processing phase is guided toward the class variable while building the model. Some common applications for classification include loan approval, medical diagnoses, email filtering etc. In unsupervised learning no labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

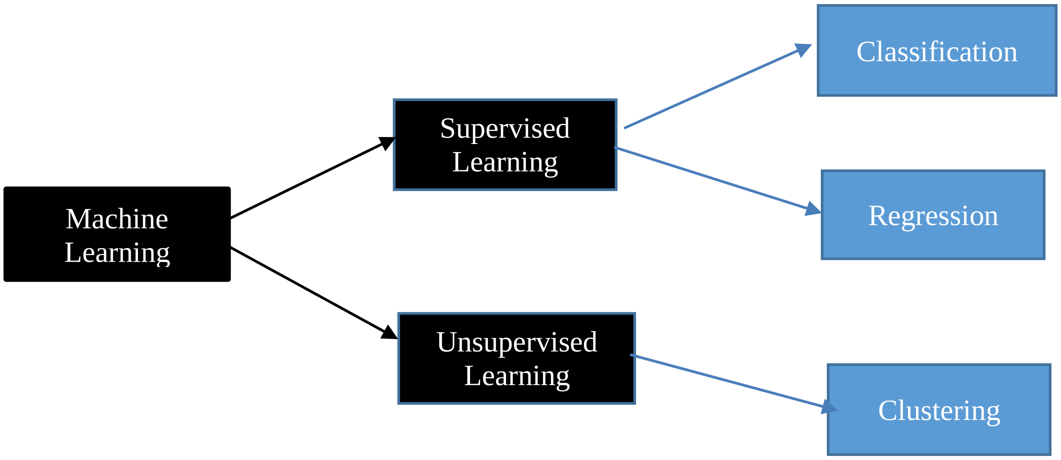


Figure1. Supervised learning versus unsupervised learning

Therefore, using machine learning for automating arrhythmia diagnosis can be very helpful.

The project aims at using different Machine learning algorithms like KNN, SVC, Random Forests and Xgboost for predicting and classifying arrhythmia into different categories. All the features which are used for classification were extracted from MATLAB. For enhancing the performance of our model K – Folds cross validation method is used. The cross validation technique can be used to compare the performance of different machine learning models on the same data set**.**

Block diagram:

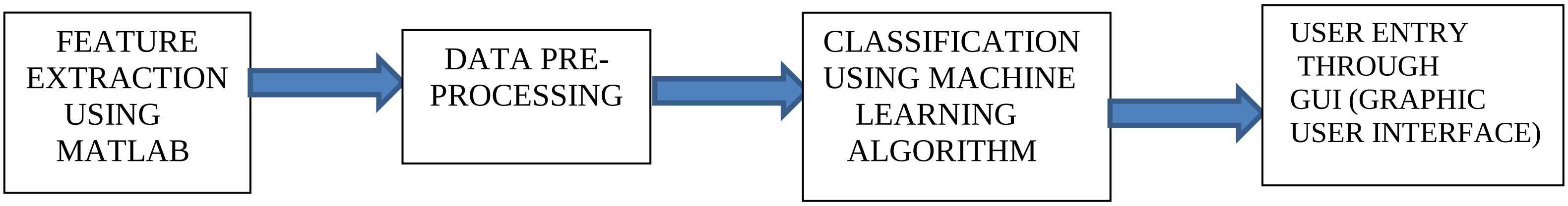
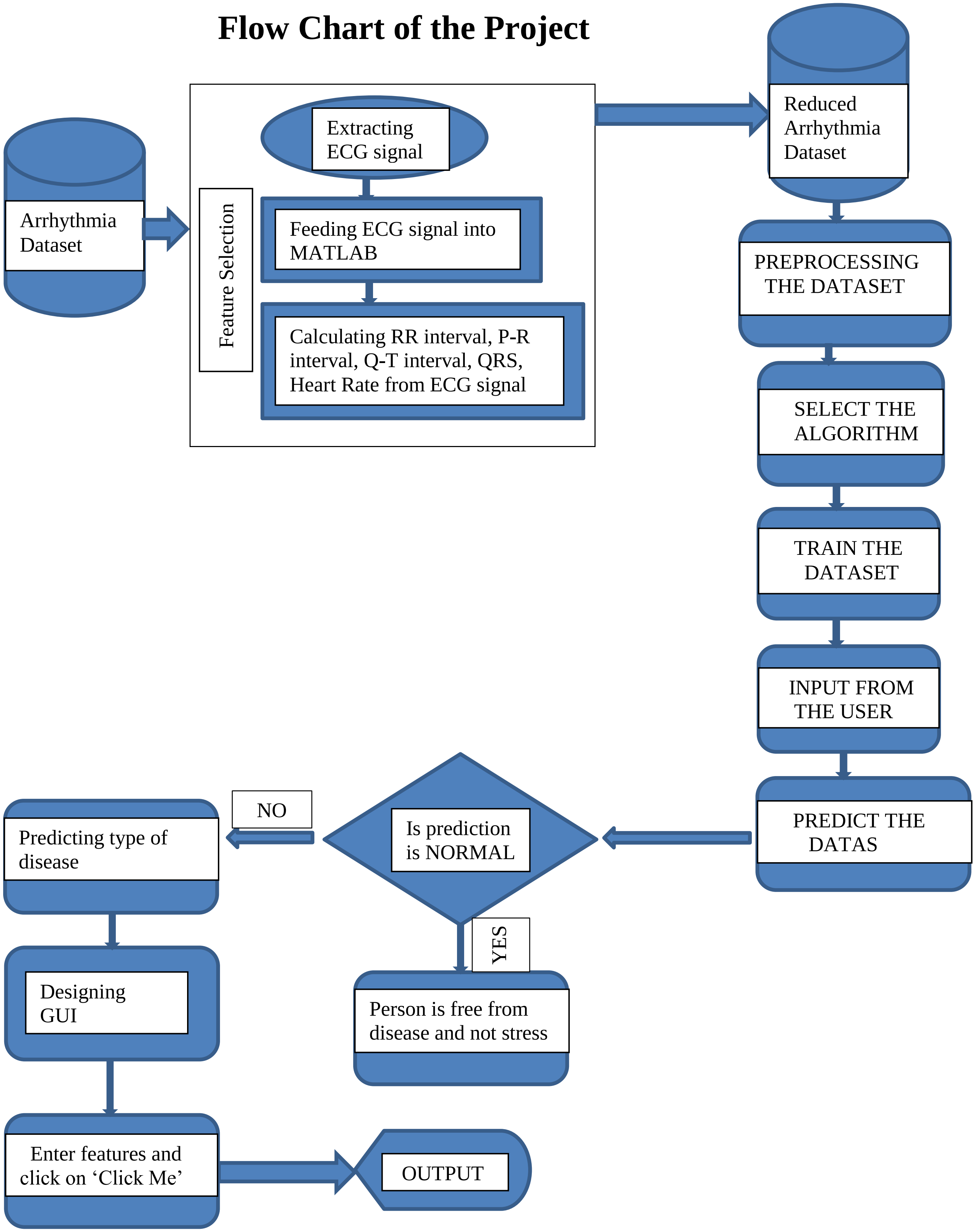

FEATURE EXTRACTION USING MATLAB
DATA PRE-PROCESSING
CLASSIFICATION USING MACHINE LEARNING ALGORITHM
USER ENTRY THROUGH GUI (GRAPHIC USER INTERFACE)

Figure2. Block Diagram for disease classification.

Flow Chart of the Project


Arrhythmia Dataset
Feature Selection
Extracting ECG signal
Feeding ECG signal into MATLAB
Calculating RR interval, P-R interval, Q-T interval, QRS, Heart Rate from ECG signal
Reduced Arrhythmia Dataset
PREPROCESSING THE DATASET
SELECT THE ALGORITHM
TRAIN THE DATASET
INPUT FROM THE USER
PREDICT THE DATAS
Is prediction is NORMAL
NO
YES
Predicting type of disease
Person is free from disease and not stress
Designing GUI
Enter features and click on 'Click Me'
OUTPUT

Detailing of the project work:

In this project we have designed a model which is predicting whether a person having disease or Not and classifying it into different stress related arrhythmia disease.

For this, we have performed following steps:

Signal Acquisition: -

Electrocardiogram (ECG) signal is the graphical representation of the electrical activity of the heart over a period of time which is recorded by the electrodes connected to the body either using the three leads or twelve leads attached to the surface of the body. In this project we take 2483 samples of the ECG of a human body. It is then divided into group of 100 samples in a set which makes total of 25 set of samples. These samples then went into the signal processing module. The publicly available database has been used in the study. These databases were recorded in the hospital using the ECG signal acquisition module.

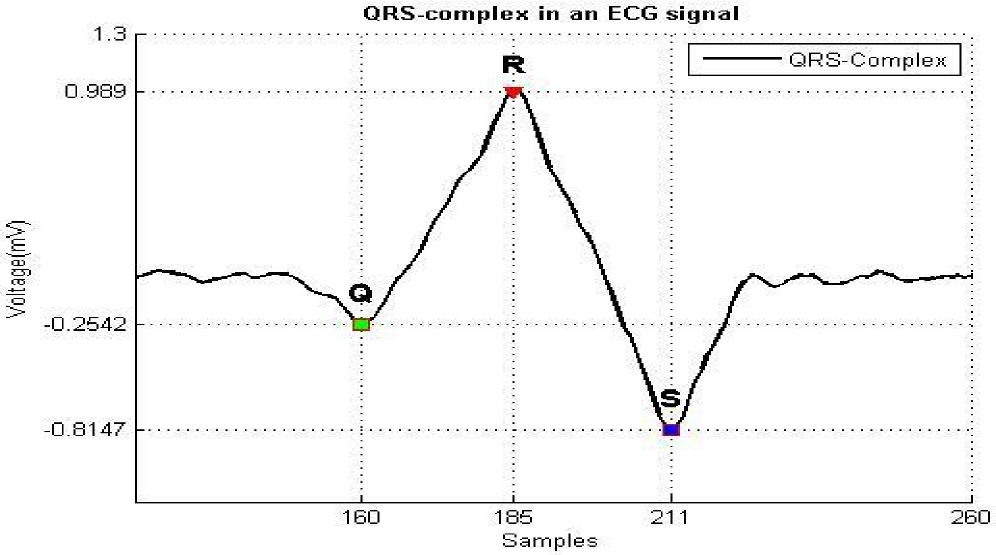
Features Extraction:-

In this case the parameters which have the potential to discriminate between various classes are calculated from the ECG signal and are further used for classification of the signals. They may be time domain, frequency domain, and statistical based parameters. The features that have been used in this study are extracted using MTALAB (R2018a).

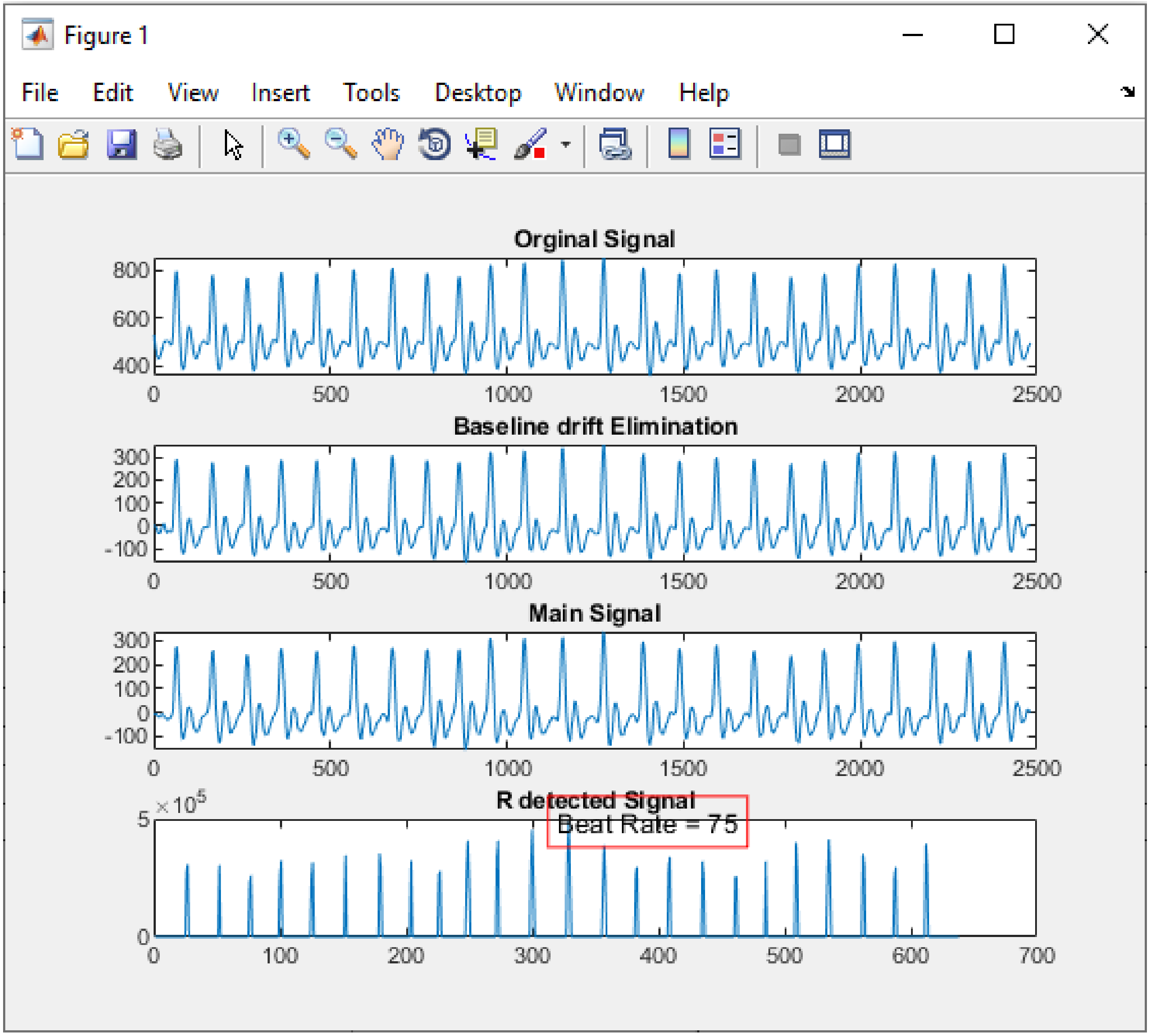
There are some features extracted. These are:-

1.) QRS-complex:-

QRS-complex is the most prominent repeating peak in the ECG signal. The QRS complex refers to the polarization and Depolarization of the right and left ventricles of the human heart. It can be used to determine the patient's cardiac rate or predict abnormalities in the heart function. The following figure shows the graph QRS complex in the ECG signal



2.) Heart Rate:- This is one of the most important feature used in classification of disease.



Heart Rate extracted from ECG Signal using MATLAB

DATASET:-

The dataset for the project is taken from the UCI Machine Learning Repository https://archive. ics.uci.edu/ml/datasets/Arrhythmia (1 csv file, 1 information file).There are (452) rows, each representing medical record of a different patient. There are 279 features like age, weight and patient's ECG related data. But we have selected only that features that we had extracted through MATLAB and they are heart rate, QRS duration (in ms), P-R interval (in ms), Q-T interval (in ms), T interval (in ms), QRS and many more features. The data set is labelled with 16 different classes. Classes 2 to 15 correspond to different types of arrhythmia. Class 1 corresponds to normal ECG with no arrhythmia and Class 16 refers to unlabelled patient.

Table 1: Class Distribution in UCI arrhythmia dataset

Class	Number of instances
Normal	245
Ischemic changes	44
Old Anterior Myocardial Infarction	15
Old Inferior Myocardial Infarction	15
Sins tachycardia	13
Sinus bacdycardia	25
Ventricular Premature Contraction	3
Supraventricular Premature Contraction	2
Left bundle branch block	9
Right bundle branch block	50
First degree AtrioVentricularblock	0
Second degree AtrioVentricularblock	0
Third degree AtrioVentricularblock	0
Left ventricular hypertrophy	4
Atrial Fibrillation	5
Others	22

Data Pre-processing:-

- Importing the dataset
- Missing Data: Replacing the missing data by mean or by medium
- Splitting the dataset into the Training set and Test set: - The training set is a subset of your data on which your model will learn how to predict the dependent variable with the independent variables. The test set is the complimentary subset from the training set, on which you will evaluate your model to see if it manages to predict correctly the dependent variable with the independent variables.
- Feature Scaling:-
  Feature scaling in machine learning is one of the most critical steps during the pre-processing of data before creating a machine learning model. Scaling can make a difference between a weak machine learning model and a better one. The most common techniques of feature scaling are Normalization and Standardization**.**

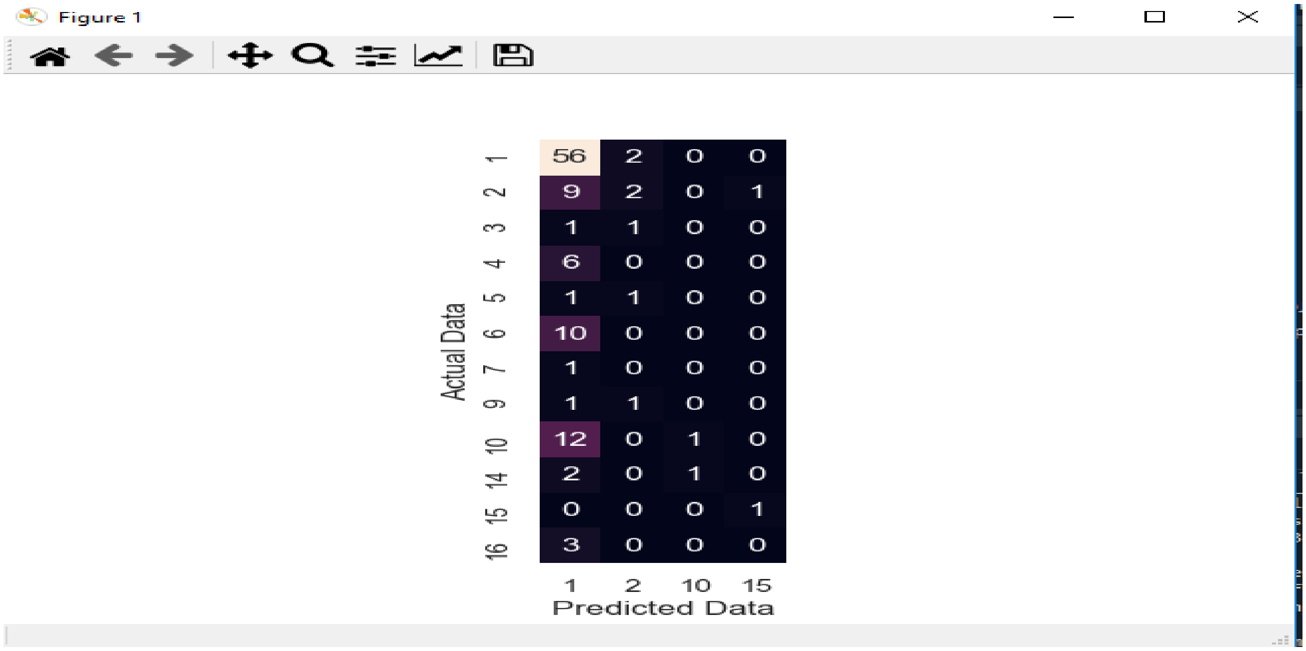
$\mathbf{X}_{\text{Standardization}} = \mathbf{(x - mean(x)) \div (Standard\ deviation(x))}$ $\mathbf{X}_{\text{Normalization}} = \mathbf{(x\text{-}min(x)) \div (max(x)\text{-}min(x))}$

- Train the model:-

We trained the model using different algorithm such as SVM (Support Vector Machine), KNN (K-nearest neighbor), and Random Forest. The different results obtained by these algorithms.

SVM: -

SVMs are the most popular algorithm for classification in machine learning algorithms**.** It is a type of supervised machine learning algorithm that provides analysis of data for classification and regression analysis. The aim of using SVM is to correctly classify unseen data. We used Kernel SVM for classification of diseases. An SVM kernel basically adds more dimensions to a low dimensional space to make it easier to segregate the data. It converts the inseparable problem to separable problems by adding more dimensions using the kernel trick. A support vector machine is implemented in practice by a kernel. The kernel trick helps to make a more accurate classifier. SVM contain a class called SVC which is used for classification and dataset was split into 75% - 25% between train and test respectively.



	1	2	10	15
1	56	2	0	0
2	9	2	0	1
3	1	1	0	0
4	6	0	0	0
5	1	1	0	0
6	10	0	0	0
7	1	0	0	0
9	1	1	0	0
10	12	0	1	0
14	2	0	1	0
15	0	0	0	1
16	3	0	0	0

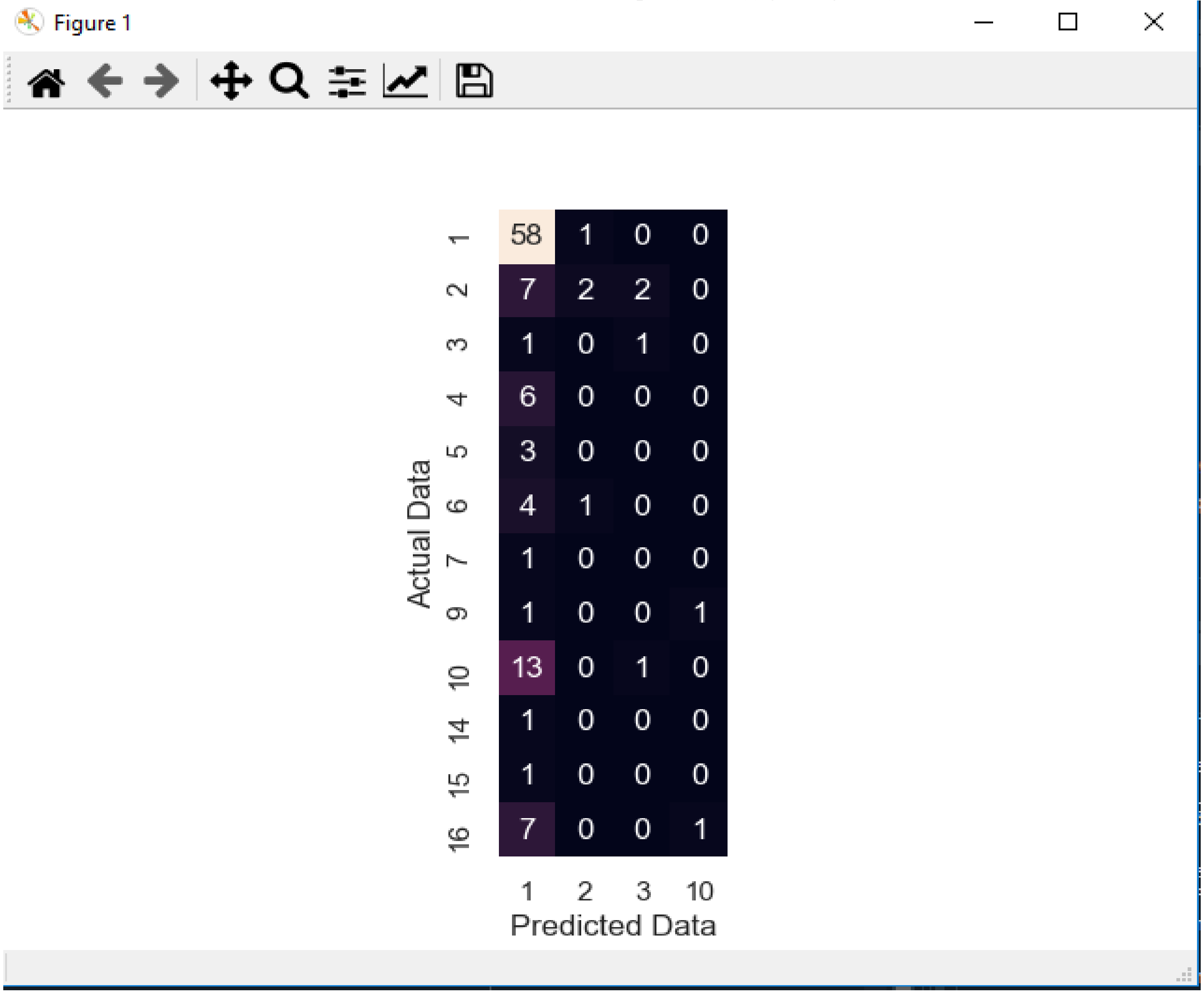
Fig6: Confusion Matrix SVM with polynomial degree 2 kernel

KNN:-

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm at the training phase just stores the dataset and when it gets new data, and then it classifies that data into a category that is much similar to the new data. To determine which of the K instances in the training dataset are most similar to a new input a distance measure is used. For real-valued input variables, the most popular distance measure is Euclidean distance.

Euclidean distance is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (xi) across all input attributes j

Euclidean Distance(x, xi) = sqrt (sum ((xj – xij) ^2))

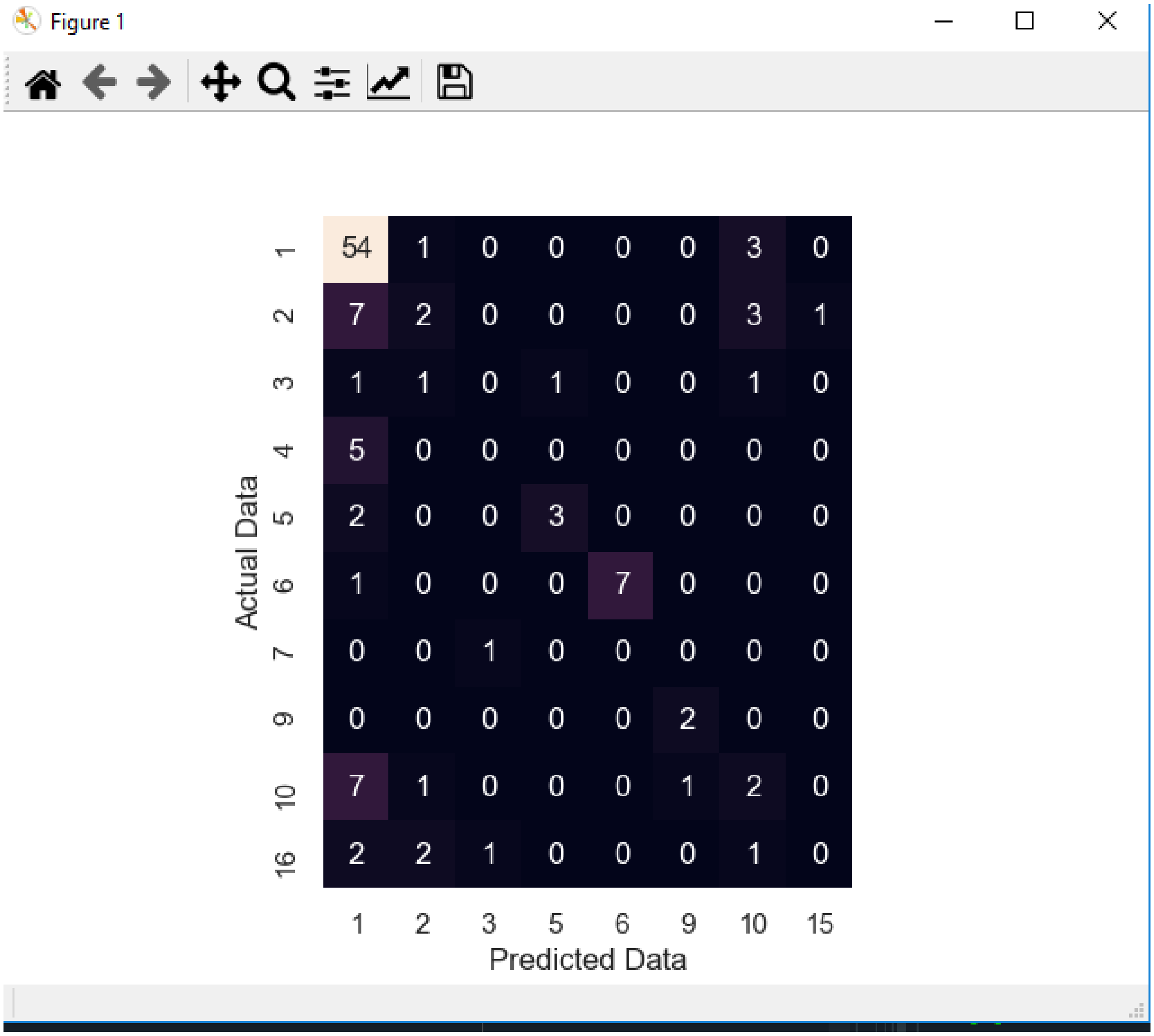


	1	2	3	10
1	58	1	0	0
2	7	2	2	0
3	1	0	1	0
4	6	0	0	0
5	3	0	0	0
6	4	1	0	0
7	1	0	0	0
9	1	0	0	1
10	13	0	1	0
14	1	0	0	0
15	1	0	0	0
16	7	0	0	1

Fig7: Confusion Matrix for KNN

Random Forest:-

It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. A simple decision tree gives good predictions when there is a huge number of a predictor variable like in this data set. Early methods to construct decision trees were unstable with small perturbations in data resulting in large changes in predictions. Random forests is an ensemble classifier that consists of many decision trees and outputs the class that is the mode of the classes output by individual trees. In this way, an RF ensemble classifier performs better than a single tree. The dataset was split as train-test 80-20 % respectively.



	1	2	3	5	6	9	10	15
1	54	1	0	0	0	0	3	0
2	7	2	0	0	0	0	3	1
3	1	1	0	1	0	0	1	0
4	5	0	0	0	0	0	0	0
5	2	0	0	3	0	0	0	0
6	1	0	0	0	7	0	0	0
7	0	0	1	0	0	0	0	0
9	0	0	0	0	0	2	0	0
10	7	1	0	0	0	1	2	0
16	2	2	1	0	0	0	1	0

Fig8: Confusion Matrix for RANDOM FOREST

GUI For User Entry:-

For making our project more interactive we develop GUI for our project .We used tkinter library of python. Python offers multiple options for developing GUI (Graphical User Interface). Out of all the GUI methods, tkinter is the most commonly used method. It is a standard Python interface to the Tk GUI toolkit shipped with Python. Python with tkinter is the fastest and easiest way to create the GUI applications. User has to enter all the features in text box and to see the result click on “Click Me” button. As a result we will see the name of the disease.

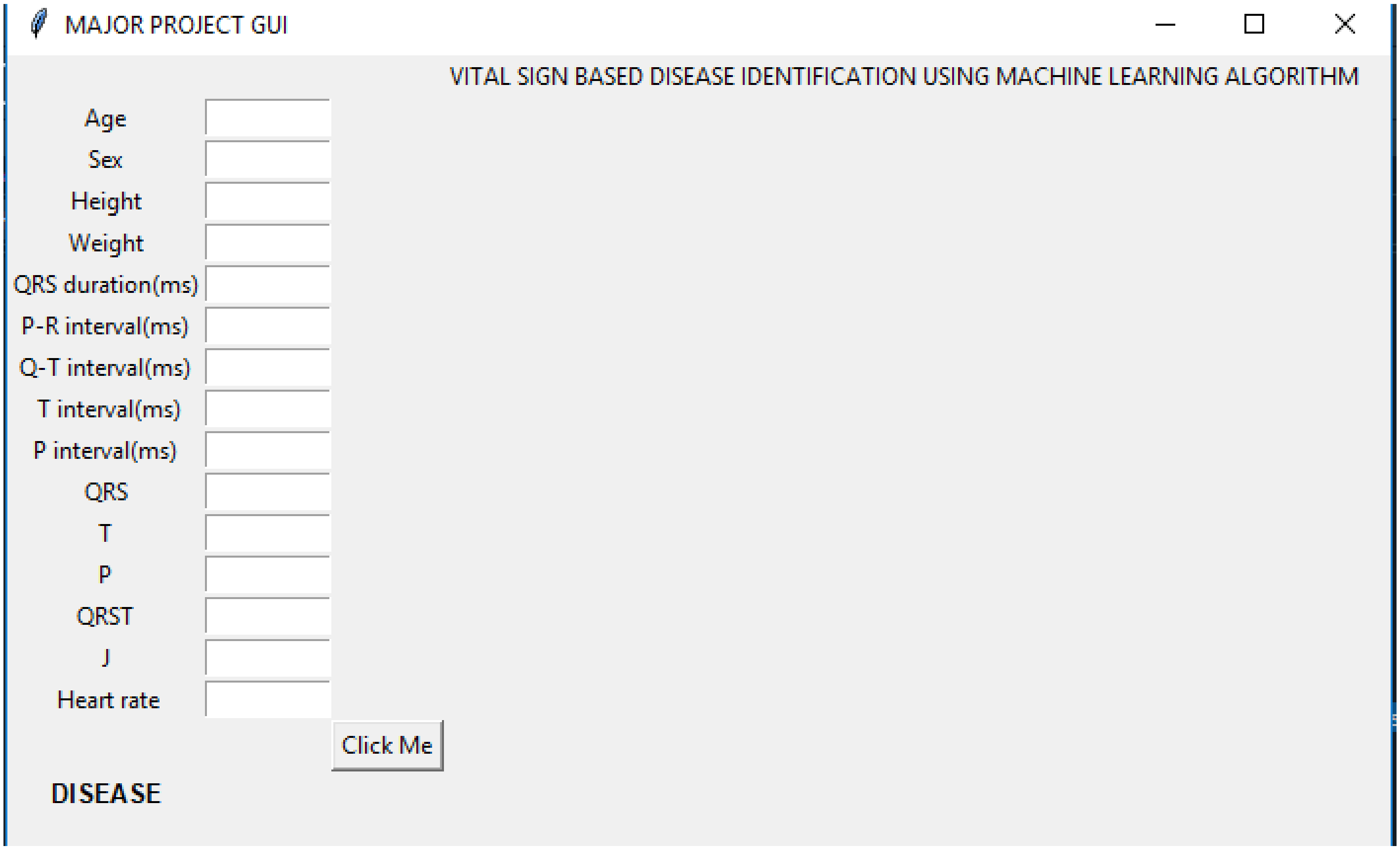
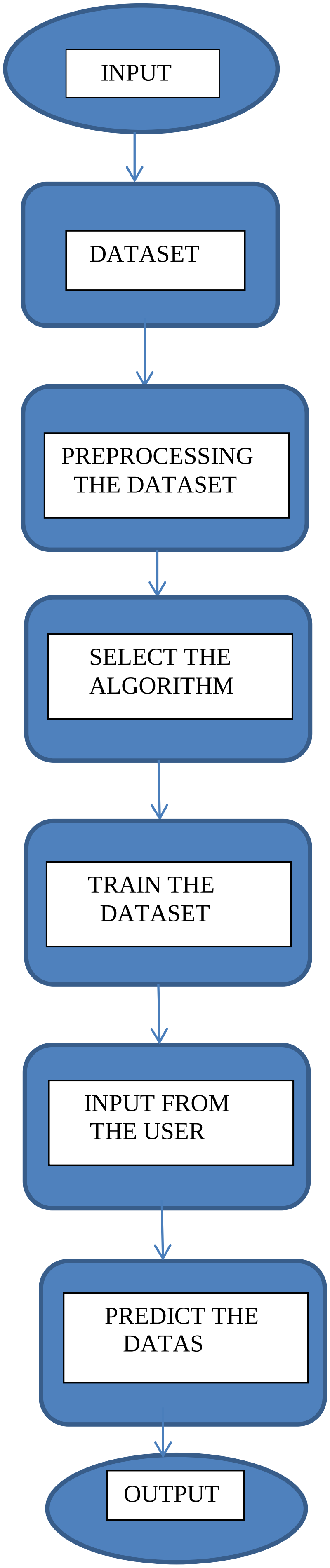


Fig 9: GUI for Project

FLOW CHART OF THE PROGRAMMING CODE:



Result Set:

Graph generated using Random Forest Algorithm for classification

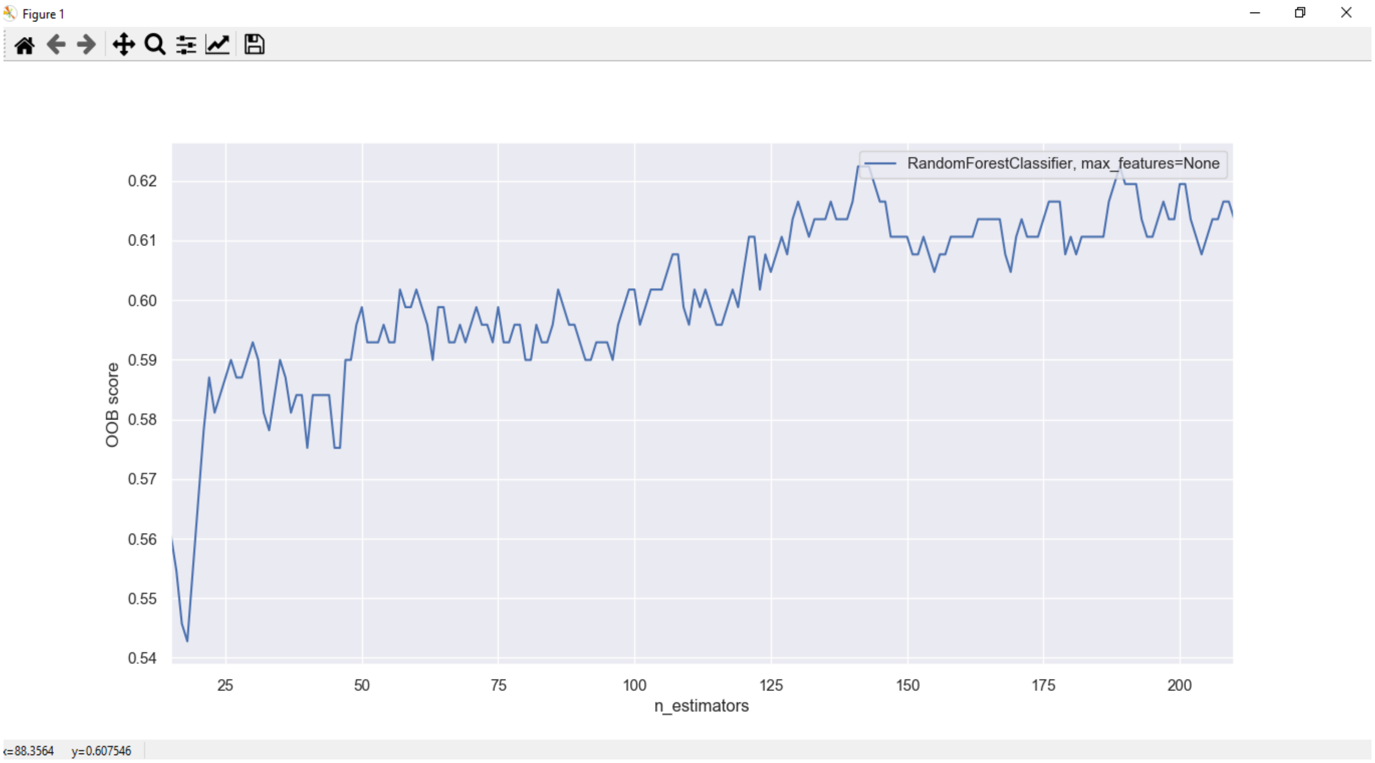
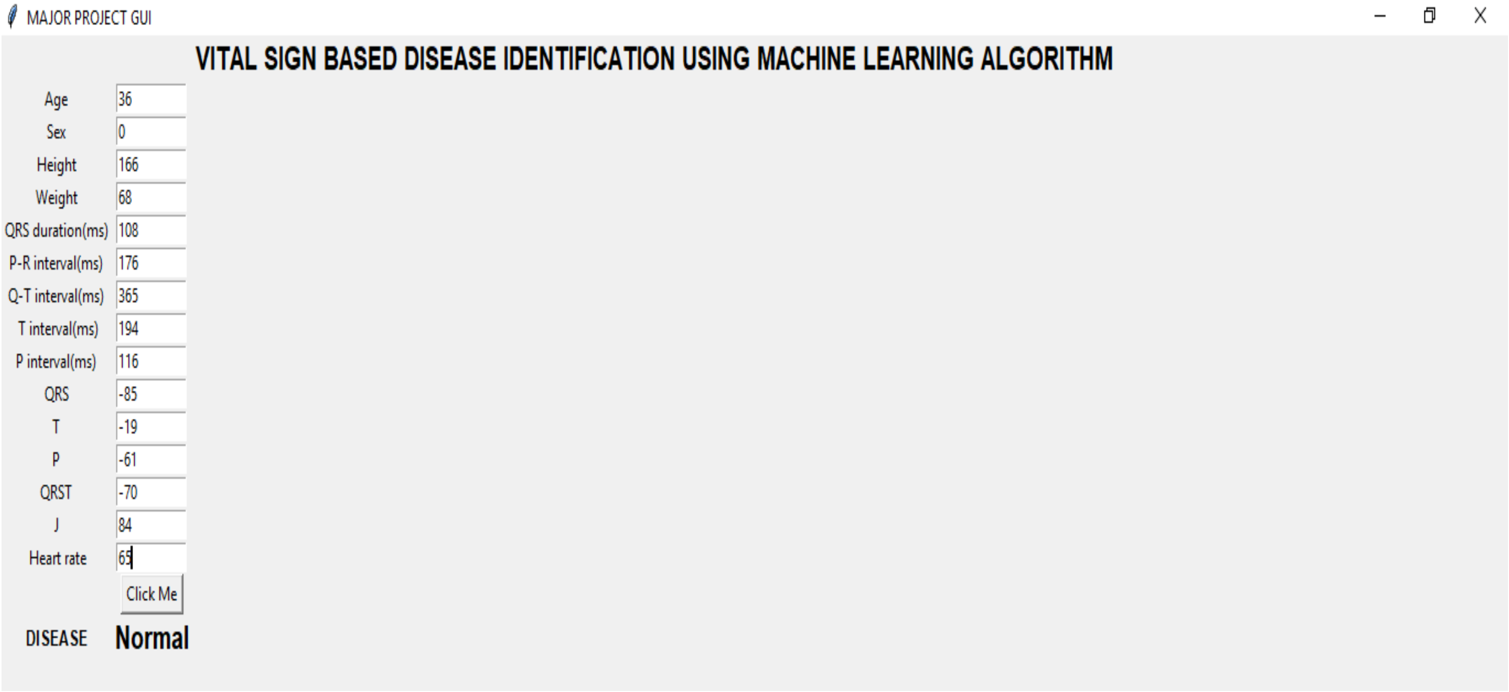


Fig10: Graph between OOB Score vs. n_estimators

The graph showing the random forest accuracy with fifteen features. Here n_estimators means no. of decision trees and OOB score is the accuracy. By using this graph we can increase the accuracy of our model. Through this graph we can interpret the highest n_estimator value i.e. 140 then the accuracy corresponding to that n_estimator is around 62.5%.

Fig 11: GUI Result for Our Project



VITAL SIGN BASED DISEASE IDENTIFICATION USING MACHINE LEARNING ALGORITHM

Age 36

Sex 0

Height 166

Weight 68

QRS duration(ms) 108

P-R interval(ms) 176

Q-T interval(ms) 365

T interval(ms) 194

P interval(ms) 116

QRS -85

T -19

P -61

QRST -70

J 84

Heart rate 65

DISEASE Normal

By entering all the features i.e. Age, Sex(0-male,1-Female), Height, Weight, QRS duration(ms), P-R interval(ms), Q-T interval(ms), T interval(ms), P interval(ms), QRS ,T ,P ,QRST,J, Heart Rate(no. of heart beat per minute) and then clicking on Click Me button we can see our result. All the datas entered here are extracted from ecg signal through MATLAB and the result obtained here is normal.

TABLE 2. ALGORITHM ACCURACY TABLE

Algorithm For Classification	**Accuracy (in %)**
SVM(Support Vector Machine)	54
KNN(K-nearest neighbor)	55
Random Forest	70

In this project we have used Random Forest Classifier to predict the disease which has got us the highest accuracy of 70% approx.

Conclusion:

Python Scikit learn software package is used via Spider 4.1.3.

This report presents a comparative study for classification of cardiac arrhythmia diseases dataset. The research is carried out on the standard dataset taken from the University of California. Several classification algorithms were examined. Our main objective of building an expected stress related disease classification model by exploring different Machine Learning techniques has been accomplished. Indeed, we used modern Machine Learning algorithms such as KNN, Random Forest and Support Vector Machines techniques for disease predictions.

Firstly we have used SVM algorithm using SVC class but the accuracy we got was only 54%.Even we are applying k-fold cross validation for enhancing the performance of model .Then we used KNN algorithm and the accuracy we got was 55%.We tried to obtain the maximum accuracy for different values of K. Finally we applied Random Forest algorithm and got the highest accuracy i.e. 70%.

Future Work:-

A number of combinations of algorithms can be implement in the hierarchical scheme. We can also apply a new approach with random forest classification where instead of one we train two different RF classifiers, the first one provides a binary classification about whether the person has arrhythmia or not. Then we further sub-classify the instances which are predicted with arrhythmia using the second random forest classifier. We can expand this to add more levels and try it with other models. We can also feed forward neural network for classification or regression with a single layer of hidden nodes would provide better result as compared to other algorithms. Also for the Network implementation, we think bootstrapping may help improve the performance

Application Of Our Project:-

With the help of our project we can implement mobile based system design .In this, service can be accessed when needed using a mobile application, which sends ECG signals remotely using Bluetooth from a wearable sensor attached to the body of the patient. This sensor could be attached as a chest belt or a watch. On the other side; the mobile application receives the ECG signal and saves it as a time series of voltage magnitudes for five minutes. This data is send to the server to be saved, analyses through the Arrhythmia Diagnosis algorithm which diagnosis, and detect the class of heart arrhythmia disease that the patient holds. The server then sends the result to alert the registered Doctor through SMS on his phone; the doctor can view the ECG history of his patient, send a medication or request an appointment through a responsive web application.

References and Bibliography

Sample References:

1. M. Sanjeev Arulampalam, Simon Maskell, Neil Gordon, and Tim Clapp, “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking”, IEEE Transactions on Signal Processing, Vol. 50, No. 2, pp:174-188, Feb’ 2002.

2. R. N. Banavar, J. L. Speyer, “Properties of Risk-Sensitive Filters/Estimators”, IEEE Proceedings of Control Theory Application, Vol.145, No. 1, January 1998.

3. R. G. Brown, and P. Y. C. Hwang, Introduction to Random Signals and Applied Kalman Filtering with Matlab Exercises and Solutions, 3rd Edition, John Wiley & Sons, Inc, 1997.

4. Universal Description, Discovery and Integration, UDDI;http://www.uddi.org; October
5. 5, 2007.

5. International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

6.H. Altay Guvenir, Burak Acar, Gulsen Demiroz, Ayhan Cekin “A Supervised Machine Learning Algorithm for Arrhythmia Analysis.” Proceedings of the Computers in Cardiology Conference, Lund, Sweden, 1997

7. zift, Akin.“Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis.” Computers in Biology and Medicine 41.5 (2011): 265-271

8. Hall, Mark A., and Lloyd A. Smith. ”Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper.” FLAIRS conference. 1999.

9. Uyar, Asl, and Fikret Gurgen. ”Arrhythmia classification using serial fusion of support vector machines and logistic regression.” Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 2007. IDAACS 2007. 4th IEEE Workshop on. IEEE, 2007.

10. Polat, Kemal, Seral ahan, and Salih Gne. ”A new method to medical diagnosis: Artificial immune recognition system (AIRS) with fuzzy weighted preprocessingand application to ECG arrhythmia.” Expert Systems with Applications 31.2 (2006): 264- 269

12.Singh N and Singh P 2018 Engineering Vibration, Communication and Information Processing, Springer,

Singapore 478 469-480

13 Rajni R and Kaur I 2013 International Journal of Computer Applications 84 22-25

14.UCI Machine Learning Repository (2018, Nov. 28). Arrhythmia Data Set. Retrieved from http://archive.ics.uci.edu/ml/datasets/Arrhythmia.

15. Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 Journal of artificial intelligence research

16 321-357

16. Haibo H, Yang B, Edwardo A G and Shutao L 2008 " ADASYN: Adaptive synthetic sampling approach for imbalanced learning " IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence) pp. 1322-1328

17. Eduardo J S L, William R S, Guillermo C-C and David M 2016 Computer Methods and Programs in Biomedicine 127 144–164.

18.Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M and Perrot M 2011 Journal of Machin Learning Research 12 2825-2830

19.Bauer E and Kohavi R 1999 Machine Learning (Berlin: Springer) (vol 36) pp 105-139

20.World Health Organization (2018, Nov. 11). EGYPT : CORONARY HEART DISEASE. Retrieved from www.worldlifeexpectancy.com/egypt-coronary-heart-disease.

21. Mitra M and Samanta R K 2013 Procedia Technology 10 76-84

22. Umale V, Waghe S, Bhalerao D and Gound R 2016 International Research Journal of Engineering and Technology (IRJET) 3 258-262

23. Mustaqeem A, Anwar S M and Majid M 2018 Computational and Mathematical Methods in Medicine 2018 Article ID 7310496

24. Arrhythmia Disease Classification and Mobile Based System Design
Soha Samir AbdElMoneem et al 2020 J. Phys.: Conf. Ser. 1447 012014

25.Udemy Course for Machine Learning

26.Scikit learn-0.23.0

Appendix: